Time-frequency processing - Spectral properties

Tuomas Virtanen, Emmanuel Vincent, Sharon Gannot

▶ **To cite this version:**

Tuomas Virtanen, Emmanuel Vincent, Sharon Gannot. Time-frequency processing - Spectral properties. Emmanuel Vincent; Tuomas Virtanen; Sharon Gannot. Audio source separation and speech enhancement, Wiley, 2018, 978-1-119-27989-1. hal-01881426

HAL Id: hal-01881426
https://inria.hal.science/hal-01881426

Submitted on 25 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

2
Time-frequency processing – Spectral properties

Tuomas Virtanen, Emmanuel Vincent, and Sharon Gannot

Many audio signal processing algorithms typically do not operate on raw time-domain audio signals, but rather on *time-frequency representations*. A raw audio signal encodes the amplitude of a sound as a function of time. Its Fourier spectrum represents it as a function of frequency, but does not represent variations over time. A time-frequency representation presents the amplitude of a sound as a function of both time and frequency, and is able to jointly account for its temporal and spectral characteristics (Gröchenig, 2001).

Time-frequency representations are appropriate for three reasons in our context. First, separation and enhancement often require modeling the structure of sound sources. Natural sound sources have a prominent structure both in time and frequency, which can be easily modeled in the time-frequency domain. Second, the sound sources are often mixed convolutively, and this convolutive mixing process can be approximated with simpler operations in the time-frequency domain. Third natural sounds are more sparsely distributed and overlap less with each other in the time-frequency domain than in the time or frequency domain, which facilitates their separation.

In this chapter we introduce the most common time-frequency representations used for source separation and speech enhancement. Section 2.1 describes the procedure for calculating a time-frequency representation and converting it back to the time domain, using the *short-time Fourier transform* (STFT) as an example. It also presents other common time-frequency representations and their relevance for separation and enhancement. Section 2.2 discusses the properties of sound sources in the time-frequency domain, including sparsity, disjointness, and more complex structures such as harmonicity. Section 2.3 explains how to achieve separation by time-varying filtering in the time-frequency domain. We summarize the main concepts and provide links to other chapters and more advanced topics in Section 2.4.

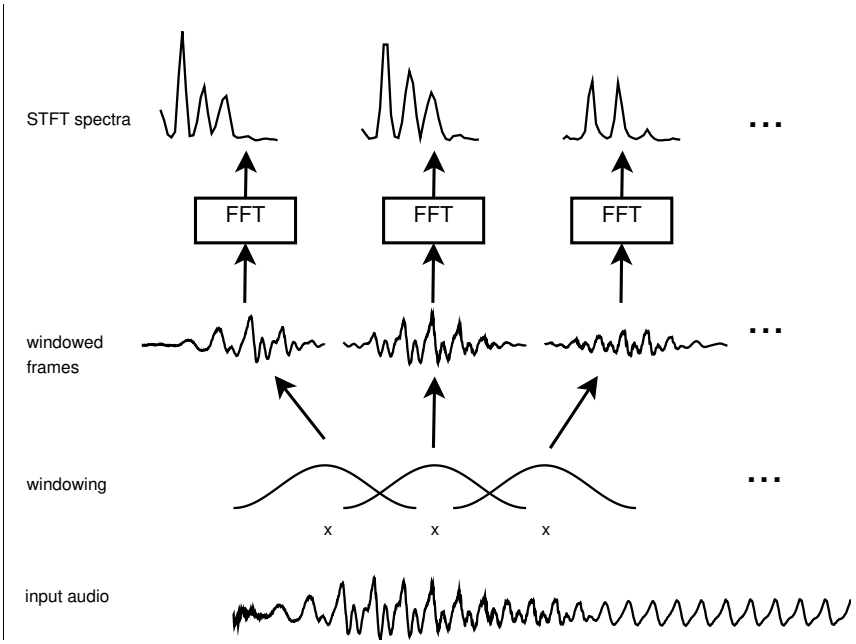

Figure 2.1 STFT analysis.

2.1 Time-frequency analysis and synthesis

In order to operate in the time-frequency domain, there is a need for analysis methods that convert a time-domain signal to the time-frequency domain, and synthesis methods that convert the resulting time-frequency representation back to the time domain after separation or enhancement. For simplicity, we consider the case of a single-channel signal ($I = 1$) and omit the channel index $i = 1$. In the case of multichannel signals, the time-frequency representation is simply obtained by applying the same procedure individually to each channel.

2.1.1 STFT analysis

Our first example of time-frequency representation is the STFT. It is the most commonly used time-frequency representation for audio source separation and speech enhancement due to its simplicity and low computational complexity in comparison to the available alternatives. Figure 2.1 illustrates the process of segmenting and windowing an audio signal into frames, and calculating the *discrete Fourier transform* (DFT) spectrum in each frame. For visualization the figure uses the magnitude spectrum $|x(n, f)|$ only, and does not present the phase spectrum $\angle x(n, f)$.

The first step in the STFT analysis (Allen, 1977) is the segmentation of the input signal into fixed-length *frames*. Typical frame lengths in audio processing vary between 10 and 120 ms. Frames are usually overlapping — most commonly by 50% or 75%. After segmentation, each frame is multiplied elementwise by a window function. The segmented and windowed signal $x(n,t)$ in frame $n \in \{0,\dots,N-1\}$ can be defined as

$$x(n,t) = x(t + t_0 + nM)h_\mathrm{a}(t), \quad t \in \{0,\dots,T-1\} \tag{2.1}$$

where N is the number of time frames, T is the number of samples in a frame, t_0 positions the first sample of the first frame, M is the hop size between adjacent frames in samples, and $h_\mathrm{a}(t)$ is the *analysis window*.

Windowing with an appropriate analysis window alleviates the spectral leakage which takes place when the DFT is applied to short frames. Spectral leakage means that energy from one frequency bin leaks to neighboring bins: even when the input frame consists of only one sinusoid, the resulting spectrum is nonzero in other bins too. The shorter the frame, the stronger the leakage. Mathematically, this can be modeled as the convolution of signal spectrum with the DFT of the window function.

For practical implementation purposes, window functions have a limited support, i.e., their values are zero outside the interval $[0, T-1]$. Typical window functions such as sine, Hamming, Hann, or Kaiser-Bessel are nonnegative, symmetric, and bell-shaped, so that the value of the window is largest at the center, and decays towards the frame boundaries. The choice of the window function is not critical, as long as a window with reasonable spectral characteristics (sufficiently narrow main lobe, and low level of sidelobes) is used. The choice of the frame length is more important as discussed later in Section 2.1.3.

After windowing, the DFT of each windowed frame is taken, resulting in complex-valued STFT coefficients

$$x(n,f) = \sum_{t=0}^{T-1} x(n,t)e^{-2\jmath\pi tf/F}, \quad f \in \{0,\dots,F-1\} \tag{2.2}$$

where F is the number of frequency bins, f is the discrete *frequency bin*, and $\jmath$ is the imaginary unit. Typically, $F = T$. We can also set F larger than the frame length T by zero-padding $x(n,t)$ by adding a desired number of zero entries $x(n,t) = 0$, $t \in \{T,\dots,F-1\}$, to the end of the frame.

We denote the frequency in Hz associated to the positive frequency bins $f \in \{0,\dots,\lceil F/2 \rceil\}$ as

$$\nu_f = \frac{f}{F} f_\mathrm{s} \tag{2.3}$$

with f_s the sampling frequency. The STFT coefficients for $f \in \{\lfloor F/2 \rfloor + 1,\dots,F-1\}$ are complex conjugate of those for $f \in \{\lceil F/2 \rceil - 1,\dots,1\}$ and are called negative frequency bins. In the following chapters, the negative frequency bins are often implicitly discarded, nevertheless equations are always written in terms of all

frequency bins $f \in \{0, \ldots, F-1\}$ for conciseness. Each term $e^{-2\jmath\pi tf/F}$ is a complex exponential with frequency ν_f, thus the DFT calculates the dot product between the windowed frame and complex basis functions with different frequencies.

The STFT has several useful properties for separation and enhancement:

- The frequency scale ν_f is a linear function of the frequency bin index f.
- The resulting complex-valued STFT spectrum allows to easily treat the *phase* $\angle x(n,f)$ and the *magnitude* $|x(n,f)|$ or the *power* $|x(n,f)|^2$ separately.
- The DFT can be efficiently calculated using the fast Fourier transform.
- The DFT is simple to invert, which will be discussed in the next section.

2.1.2 STFT synthesis

Source separation and speech enhancement methods result in an estimate $\widehat{c}(n,f)$ or $\widehat{s}(n,f)$ of the target source in the STFT domain. This STFT representation is then transformed back to the time domain, at least if the signals are to be listened to. Note that we omit the source index j for conciseness.

In the STFT synthesis process, the individual STFT frames are first converted to the time domain using the inverse DFT, i.e.,

$$\widehat{c}(n,t) = \frac{1}{F}\sum_{f=0}^{F-1} \widehat{c}(n,f)e^{2\jmath\pi tf/F}, \quad t \in \{0,\ldots,T-1\}. \tag{2.4}$$

The inverse DFT can also be efficiently calculated.

The STFT domain filtering used to estimate the target source STFT coefficients may introduce artifacts that affect all time samples in a given frame. These artifacts are typically most audible at the frame boundaries, and therefore the frames are again windowed by a *synthesis window* $h_{\mathrm{s}}(t)$ as $\widehat{c}(n,t)h_{\mathrm{s}}(t)$. The synthesis windows are also usually bell-shaped, attenuating the artifacts at the frame boundaries.

Overlapping frames are then summed to obtain the entire time domain signal $\widehat{c}(t)$, as illustrated in Fig. 2.2. Together with synthesis windowing, this operation can be written as

$$\widehat{c}(t) = \sum_{n=0}^{N-1} \widehat{c}(n, t-t_0-nM)h_{\mathrm{s}}(t-t_0-nM). \tag{2.5}$$

The above procedure is referred to as *weighted overlap-add* (Crochiere, 1980). It modifies the original overlap-add procedure of Allen (1977) by using synthesis windows to avoid artifacts at the frame boundaries. Even though in the above formula the summation extends over all time frames n, with practical window functions $h_{\mathrm{s}}(t)$ that are zero outside the interval $[0, T-1]$, only those terms for which $h_{\mathrm{s}}(t-t_0-nM) \neq 0$ need to be included in the summation.

The analysis and synthesis windows are typically chosen to satisfy the so-called *perfect reconstruction* property: when the STFT representation is not modified, i.e.,

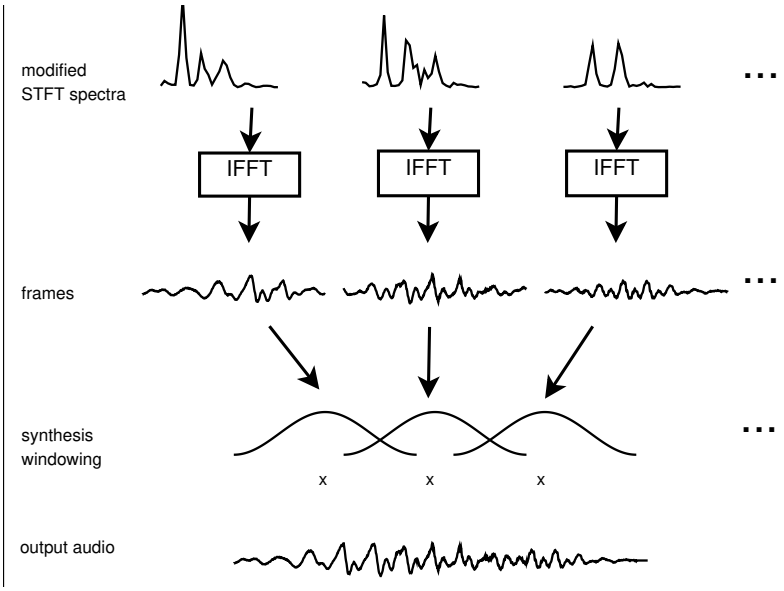

Figure 2.2 STFT synthesis.

$\widehat{c}(n, f) = x(n, f)$, the entire analysis-synthesis procedure needs to return the original time-domain signal $\widehat{c}(t) = x(t)$. Since each frame is multiplied by both the analysis and synthesis windows, perfect reconstruction is achieved if and only if condition[1)] $\sum_{n=0}^{N-1} h_{\mathrm{a}}(t - t_0 - nM)h_{\mathrm{s}}(t - t_0 - nM) = 1$ is satisfied for all t. A commonly used analysis window is the Hamming window (Harris, 1978), which gives perfect reconstruction when no synthesis window is used (i.e., $h_{\mathrm{s}}(t) = 1$). Any such analysis window that gives perfect reconstruction without a synthesis window can be transformed to an analysis-synthesis window pair by taking a square root of it, since effectively the same window becomes used twice, which cancels the square root operation.

2.1.3 Time and frequency resolution

Two basic properties of a time-frequency representation are its time and frequency resolution. In general, the time resolution is characterized by the window length and the hop size between adjacent windows, and the frequency resolution is characterized by the center frequencies and the bandwidths of individual frequency bins.

In the case of the STFT, the window length T is fixed over time and the hop size M can be freely chosen, as long as the perfect reconstruction condition is satisfied. The

1) This expression simplified from the original by Portnoff (1980) assumes that the analysis and the synthesis windows have equal lengths.

frequency scale ν_f is linear so the difference between two adjacent center frequencies $\nu_{f+1} - \nu_f = f_s/F$ is constant. The bandwidth of each frequency bin depends on the used analysis window, but is always fixed over frequency and inversely proportional to the window length T. The bandwidth in which the response of a bin falls by 6 dB is on the order of $2f_s/T$ Hz for typical window functions.

From the above we can see that the frequency resolution and the time resolution are inversely proportional to each other. When the time resolution is high, the frequency resolution is low, and vice-versa. It is possible to decrease the frequency difference between adjacent frequency bins by increasing the number of frequency bins F in (2.2). This operation called *zero padding* is simply achieved by concatenating a sequence of zeros after each windowed frame before calculating the DFT. It effectively results in interpolating the STFT coefficients between frequency bins, but does not affect the bandwidth of the bins, nor the capability of the representation to resolve frequency components that are close to each other.

Due to its impact on time and frequency resolution, the choice of the window length T is critical. Most of the methods discussed in this book benefit from time-frequency representations where sources to be separated exhibit little overlap in the STFT domain, and therefore the window length should depend on how stationary the sources are (see Section 2.2). Methods using multiple channels and dealing with convolutive mixtures benefit from window lengths longer from the impulse response from source to microphone, so that the convolutive mixing process is well modeled (see Section 3.4.1). In the case of separation by oracle binary masks, Vincent *et al.* (2007, fig. 5) found that a window length on the order of 50 ms (e.g., $T = 1024$ at $f_s = 16$ kHz) is suitable for speech separation, and a longer window length (e.g., $T = 4096$ at $f_s = 44.1$ kHz) for music, when the performance was measured by the signal-to-distortion ratio (SDR). For other objective evaluations of preferred window shape, window size, hop size, and zero padding see Araki *et al.* (2003) and Yılmaz and Rickard (2004).

2.1.4 Alternative time-frequency representations

Alternatively to the STFT, many other time-frequency representations can be used for source separation and speech enhancement. Adaptive representations (Mallat, 1999; ISO, 2005) whose time and/or frequency resolution are automatically tuned to the signal to be processed have achieved limited success (Nesbit *et al.*, 2009). We describe below a number of time-frequency representations that differ from the STFT by the use of a fixed, nonlinear frequency scale. These representations can be either derived from the STFT or computed via a filterbank.

2.1.4.1 Nonlinear frequency scales

The *Mel* scale (Stevens *et al.*, 1937; Makhoul and Cosell, 1976) and the *equivalent rectangular bandwidth* (ERB) scale (Glasberg and Moore, 1990) are two nonlinear

frequency scales motivated by the human auditory system[2]. The Mel scale is popular in speech processing, while the ERB scale is widely used in computational methods inspired by auditory scene analysis. A given frequency in Mel or ERB corresponds to the following frequency in Hz:

$$\nu(\text{Hz}) = 700 \times (e^{\nu(\text{Mel})/1127} - 1) \tag{2.6}$$

$$\nu(\text{Hz}) = 229 \times (e^{\nu(\text{ERB})/9.26} - 1). \tag{2.7}$$

If frequency bins or filterbank channels are linearly spaced on the Mel scale according to $\nu_f(\text{Mel}) = \frac{f}{F-1}\nu_{\max}(\text{Mel})$, $f \in \{0, \ldots, F-1\}$, where $\nu_{\max}(\text{Mel})$ is the maximum frequency in Mel, then their center frequencies $\nu_f(\text{Hz})$ in Hz are approximately linearly spaced below 700 Hz and logarithmically spaced above that frequency. The same property holds for the ERB scale, except that the change from linear to logarithmic behavior occurs at 229 Hz. The logarithmic scale (Brown, 1991; Schörkhuber and Klapuri, 2010)

$$\nu_f(\text{Hz}) = \nu_{\min}(\text{Hz}) \times 2^{f/F_{\text{oct}}} \tag{2.8}$$

with $\nu_{\min}(\text{Hz})$ the lowest frequency in Hz and F_{oct} the number of frequency bins per octave, is also commonly used in music signal processing applications, since the frequencies of musical notes are distributed logarithmically. It allows easy implementation of models where change in pitch corresponds to translating the spectrum in log-frequency.

When building a time-frequency representation from the logarithmic scale (2.8), the bandwidth of each frequency bin is generally chosen so that it is proportional to the center frequency, a property known as *constant-Q* (Brown, 1991). More generally, for any nonlinear frequency scale, the bandwidth is often set to a small multiple of the frequency difference between adjacent bins. This implies that the frequency resolution is narrower at low frequencies and broader at high frequencies. Conversely, the time resolution is narrower at high frequencies and coarser at low frequencies (when the representation is calculated using a filterbank as explained in Section 2.1.4.3, not via the STFT as explained in Section 2.1.4.2). This can be seen in Fig. 2.3, which shows example time-frequency representations calculated using the STFT and Mel scale.

These properties can be desirable for two reasons. First, the amplitude of natural sounds varies more quickly at high frequencies. Integrating it over wider bands makes the representation more stable. Second, there is typically more structure in sound at low frequencies, that is beneficial to model by using a higher frequency resolution for lower frequencies. By using a nonlinear frequency resolution, the number of frequency bins, and therefore the computational and memory cost of further processing, can in some scenarios be reduced by a factor of 4 to 8 without sacrificing the separation performance in a single-channel setting (Burred and Sikora, 2006). This is counterweighted in a multichannel setting by the fact that the narrowband model of

2) The Mel scale measures the perceived frequency ratio between pure sinusoidal signals, while the ERB scale characterizes the frequency response of peripheral auditory filters.

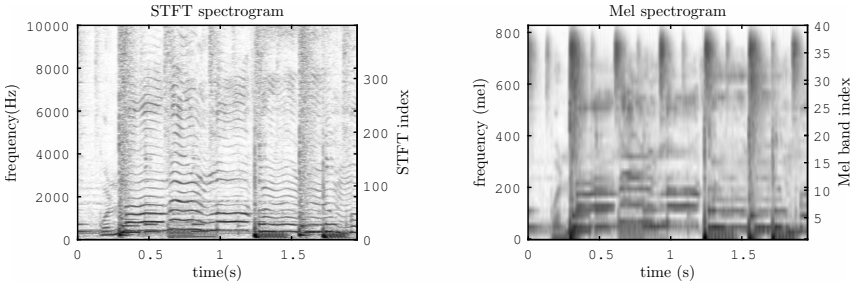

Figure 2.3 STFT and Mel spectrograms of an example music signal. High energies are illustrated with dark color and low energies with light color.

the convolutive mixing process (see Section 3.4.1) becomes invalid at high frequencies due to the increased bandwidth. Duong *et al.* (2010) showed that a full-rank model (see Section 3.4.3) is required in this case.

2.1.4.2 Computation of power spectrum via the STFT

The first way of computing a time-frequency representation on a nonlinear frequency scale is to derive it from the STFT. Even though there are methods that utilize STFT-domain processing to obtain complex spectra with nonlinear frequency scale, here we resort to methodology that estimates the power spectrum only. The resulting power spectrum cannot be inverted back to the time domain since it does not contain phase information. It can however be employed to estimate a separation filter that is then interpolated to the DFT frequency resolution and applied in the complex-valued STFT domain.

In order to distinguish the STFT and the nonlinear frequency scale representation, we momentarily index the frequency bins of the STFT by $f' \in \{0, \ldots, F'-1\}$ and the frequency bins of the nonlinear representation by $f \in \{0, \ldots, F-1\}$. The computation consists of the following steps:

1) Window the signal into frames and calculate the DFT $x(n, f')$ of each frame, similarly to the STFT analysis in Section 2.1.
2) Compute the power spectrum $|x(n, f')|^2$ in each frame.
3) Multiply this spectrum elementwise by a set of bandpass filter responses $h(f, f')$ that are linearly spaced on the chosen frequency scale.
4) Sum over f' to obtain the nonlinear spectrum $|x(n, f)|^2 = \sum_{f'=0}^{F'-1} h(f, f')|x(n, f')|^2$ for each $f \in \{0, \ldots, F-1\}$.

The Mel spectrum is usually computed using a set of triangular filter responses, as depicted in Fig. 2.4. In the multichannel case, the quantity $|x(n, f')|^2$ can be replaced by $\mathbf{x}(n, f')\mathbf{x}^H(n, f')$ which results in a quadratic time-frequency representation (Gröchenig, 2001, chap. 4) as shown by Vincent (2006). In addition to to the power spectrum, this spatial covariance matrix representation contains information about the interchannel phase and level differences (IPDs and ILDs, respectively), which are useful cues in multichannel processing.

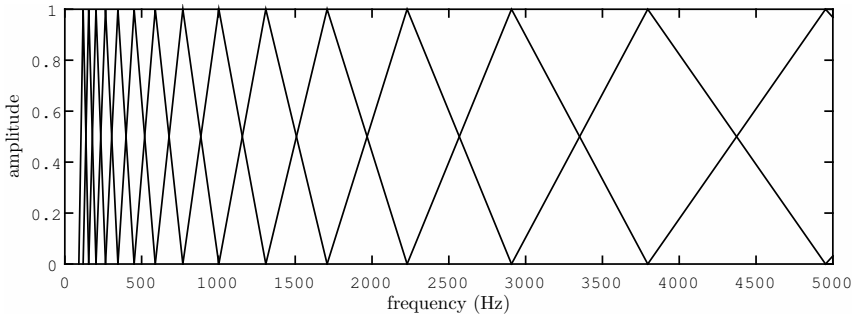

Figure 2.4 Set of triangular filter responses distributed uniformly on the Mel scale.

2.1.4.3 **Computation via a filterbank**

Alternatively, a time-frequency representation with phase information can be obtained by filterbank analysis. A filterbank consists of a set of time-domain finite impulse response (FIR) filters[3] $h_f(\tau)$, $\tau \in \{-T_f/2, \ldots, T_f/2\}$ whose center frequencies are linearly spaced on the desired scale and whose lengths T_f vary with frequency and are inversely proportional to the desired bandwidth. These filters can be generated by modulating and scaling a prototype impulse response (Burred and Sikora, 2006). The input signal $x(t)$ is convolved with each of the filters to obtain a set of complex-valued subband signals $x_f(t)$ as $x_f(t) = h_f \star x(t)$, which can then be decimated by a factor M to get $x(n, f) = x_f(nM)$. For a more detailed discussion of filterbank processing, see Vaidyanathan (1993).

The resulting representation can be approximately inverted back to the time domain by reverting the decimation operation by interpolation, convolving each subband signal by a synthesis filter, and summing the filtered signals together (Slaney *et al.*, 1994). This process of inverting the representation is approximate and causes some distortion to the signal. In many applications, the amount of distortion caused by the inversion is much smaller than the artifacts caused by the separation process, such that perfect reconstruction is not required.

Mathematically, time-frequency representations obtained via the STFT or filterbanks are very similar (see Crochiere and Rabiner (1983) for a full proof). Indeed, the STFT analysis process described in Section 2.1 can be written as $x(n, f) = \sum_{t=0}^{T-1} x(t + t_0 + nM) h_\mathrm{a}(t) e^{-2\jmath\pi tf/F}$ for each f. This corresponds to convolving the signal $x(t)$ with a time-reversed version of the FIR filter $h_\mathrm{a}(t)e^{-2\jmath\pi tf/F}$ and decimating by a factor M. Thus, STFT analysis is essentially a special case of filterbank analysis (Portnoff, 1980). Filterbanks can also be used to compute single-channel power spectra or multichannel quadratic representations by integrating the squared subband signals over time (Vincent, 2006).

3) In the general case infinite impulse response filters can also be used, but for simplicity we resort to FIR filters in this chapter.

2.2 Source properties in the time-frequency domain

Natural sound sources have several properties in the time-frequency domain which can be exploited in source separation or speech enhancement methods. In this section we discuss the most important properties of natural sound sources from this point of view.

2.2.1 Sparsity

Audio sources are sparse in the time-frequency domain, which means that only a small proportion of time-frequency bins have a significant amplitude. This is illustrated by the bottom left panel of Fig. 2.5, which shows that the vast majority of STFT coefficients of an exemplary speech signal have very low magnitude, and only a small fraction are large. This kind of distribution is termed as *sparse*.

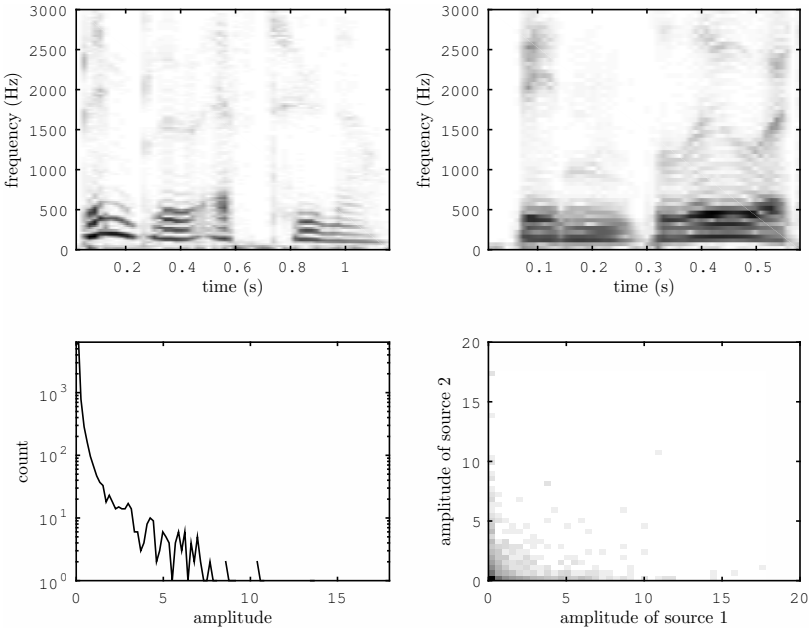

Figure 2.5 Independent sound sources are sparse in time-frequency domain. The top row illustrates the magnitude STFT spectrograms of two speech signals. The bottom left panel illustrates the histogram of the magnitude STFT coefficients of the first signal, and the bottom right panel the bivariate histogram of the coefficients of both signals.

Sparsity leads to a related phenomenon called *W-disjoint orthogonality* (Yılmaz and Rickard, 2004), which means that there is a small probability that two independent sources have significant amplitude in the same time-frequency bin. This is illus-

trated in the bottom right panel of Fig. 2.5, which shows the bivariate histogram of the STFT magnitudes of two sources. Most observed magnitude pairs are distributed along the horizontal or vertical axis, and only about 0.2% of them have a significant amplitude in the same time-frequency bin.

W-disjoint orthogonality (Yılmaz and Rickard, 2004) is the foundation for single- and multichannel classification or clustering based methods (see Chapters 7 and 12) which predict the dominant source in each time-frequency bin and for multichannel separation methods based on nongaussian source models (see Chapter 13). Furthermore, since only one source is assumed to be dominant in each time-frequency bin, the phase of the mixture is typically close to that of the dominant source. This is one motivation for assigning the phase of the mixture to the estimated dominant source (see Chapter 5). It should be noted that reverberation decreases sparsity and therefore also W-disjoint orthogonality of sources, making these separation methods less effective in reverberant spaces.

2.2.2 Structure

Natural sounds typically have structure in both time and frequency, which translates to different entries of their time-frequency representations being dependent on each other. In the simplest case, the spectrum of a highly stationary noise source changes only very little over time, making the spectrum estimation task easier (see Chapter 6). The structure can also be much more complex and present, e.g., at different time scales. Joint modeling of different time-frequency parts of a sound allows separating sources that overlap with each other in time and frequency, allowing to estimate more accurately individual source statistics even from single-channel mixtures. For example, the amplitude of a harmonic component (see discussion below about harmonic sounds) that overlaps with another source can be predicted based on the amplitudes of other harmonics of the source.

One specific type of structure is repetition over time. Natural sound sources often consist of basic units that are present multiple times over time. For example, speech consists of phonemes, syllables, and words that are used to compose utterances. Music consists of individual notes played by different instruments, that form chords, rhythmic patterns, and melodies. When processing a long audio signal, a single basic unit (e.g., phoneme, syllable, note) does not typically appear only once, but there are multiple repetitions of the unit, which are similar to each other. There exist various methods for finding repeating temporal structures (see Chapters 8, 9, 14, and 16).

Another specific type of structure within natural sound sources is *harmonicity*. Harmonic sounds have resonant modes at approximately integer multiples of the *fundamental frequency* of the sound, also called *pitch*, as shown in Fig. 2.6. Harmonic sounds include vowels in speech, notes played by most pitched musical instruments, and many animal vocalizations. In the specific case of speech, harmonic sounds are called *voiced* and other sounds are called *unvoiced*. Harmonicity has motivated source separation and speech enhancement methods that constrain the es-

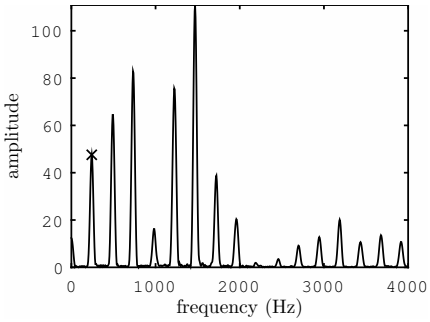

Figure 2.6 Magnitude spectrum of an exemplary harmonic sound. The fundamental frequency is marked with a cross. The other harmonics are at integer multiples of the fundamental frequency.

timated source spectra to be harmonic (see Chapter 8), and two-step methods that first track the pitch of a source over time and then predict which time-frequency bins have significant energy by considering its harmonics (see Chapter 16).

Many natural sources such as speech consist of different types of elementary components discussed above (e.g., noise-like, harmonic, or transient). Each source has slightly different characteristics that make it unique and can be used to differentiate it from other sources. Source-specific models accounting for these characteristics can be trained with appropriate machine learning algorithms such as nonnegative matrix factorization (NMF), as we shall see in, e.g., Chapters 8 and 9.

In order to take advantage of the structure of sounds discussed above, an appropriate representation of sound should be used. Time-domain representations often do not have as clear structure as time-frequency representations, since the frequency components of a source are generally not in phase lock and their phase is affected by the room impulse response. In the time-frequency domain the phase values are similarly quite stochastic and subject to different kinds of variabilities.

Therefore methods that exploit the structure of sounds in a single-channel setting often discard the phase and model the magnitude spectrum only. In a multichannel setting, the IPDs are extensively used since they bring essential information (see Chapters 3, 10, 11, 12, 14, and 18) and they exhibit less variability in comparison to the phase values as such.

2.3 Filtering in the time-frequency domain

Most source separation and speech enhancement methods apply a time-varying filter to the mixture signal. Since the source signals and the convolutive mixing process can both be modeled in the time-frequency domain, it is desirable to implement this filter in the same domain. In other words, the objective of source separation and speech enhancement methods is to estimate the target time-frequency coefficients

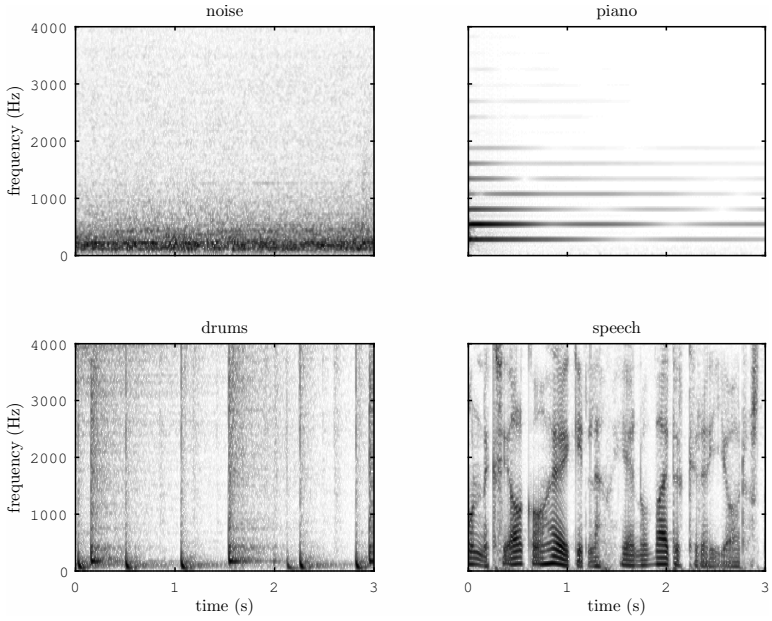

Figure 2.7 Example spectrograms of a stationary noise signal (top left), a note played by a piano (top right), a sequence of drum hits (bottom left), and a speech signal (bottom right).

$\widehat{c}(n, f)$ and $\widehat{s}(n, f)$ from the mixture coefficients $x(n, f)$.

2.3.1 Time-domain convolution as interframe and interband convolution

Let us consider how a time-domain FIR filter $w(\tau)$, $\tau \in \{0, \ldots, L-1\}$, can be implemented in the time-frequency domain. For simplicity, we consider $w(\tau)$ to be time-invariant for the moment.

Time-domain convolution can be implemented as a complex-valued multiplication in the STFT domain via either the overlap-save method (by removing the samples that underwent circular convolution) or the overlap-add method (by properly designed zero-padding of the analysis window) (Shynk, 1992). These methods are exact if both the analysis and synthesis windows are rectangular with different lengths. If the filter length L is longer than the window length, the linear convolution can still be implemented, by partitioning the filters into blocks (Servière, 2003; Soo and Pang, 1990). They have been used in the context of source separation and speech enhancement by, e.g., Gannot *et al.* (2001), Kellermann and Buchner (2003), Servière (2003), and Mirsamadi *et al.* (2012) but the use of rectangular analysis and synthesis windows severely limits their performance.

Using the conventional STFT with arbitrary analysis window instead, time-domain convolution translates into interframe and interband convolution (Gilloire and Vet-

terli, 1992; Avargel and Cohen, 2007):

$$c(n,f) = \sum_{f'=0}^{F-1} \sum_{n'=-\infty}^{\infty} w(n',f',f)x(n-n',f'). \tag{2.9}$$

This expression simply stems from the linearity of the STFT analysis operation. It is exact but computationally and memory intensive: $w(n',f',f)$ is nonzero for a few values of n' only (on the order of L/M) but for all values of f' and f. Therefore, for a given output frequency bin f, all input frequency bins f' need to be taken into account. For more discussion about this, see Chapter 19.

2.3.2 Practical approximations

The computational complexity can be reduced by noting that $w(n',f',f)$ typically decays with increasing frequency difference $|f'-f|$, where the rate of decay is governed by the window shape. A first approximation is to assume that $w(n',f',f) \approx 0$ if $f' \neq f$, which yields the *subband filtering* operation:

$$c(n,f) = \sum_{n'=-\infty}^{\infty} w(n',f)x(n-n',f). \tag{2.10}$$

In the limit when the filter length is much shorter than the analysis window length, i.e., $L \ll T$, one can further assume that $w(n',f) \approx 0$ for $n' \neq 0$, which yields the so-called *narrowband approximation*:

$$c(n,f) = w(f)x(n,f). \tag{2.11}$$

This approximation is also valid for time-frequency representations computed by filterbank analysis in the limit when $L \ll T_f$.

The majority of source separation and speech enhancement techniques employ the narrowband approximation even with a filter length equal to the frame size, namely $L = T$. This approach is convenient since, contrary to assuming that $L < T$, it does not confine the vector of filter coefficients $\mathbf{w}(n) = [w(n,0),\ldots,w(n,F-1)]^T$ to belong to an L-dimensional subspace. However, breaching the condition $L \ll T$ typically results in cyclic convolution artifacts, namely wrapping of the frames due to the filter application. These cyclic convolution effects can be alleviated by applying frequency-domain smoothing to the frequency response of the filter, tapered analysis and synthesis windows.

It should be noted that in most source separation and speech enhancement methods, time-varying filter coefficients $w(n,f)$ are used, since the sources to be separated are nonstationary.

2.4 Summary

In this chapter, we showed how a time-domain signal can be transformed to the time-frequency domain and back to the time domain, and how the time and frequency resolution of this transform can be controlled. In addition we discussed how the time-frequency coefficients of the target source can be approximately obtained by narrowband filtering in the time-frequency domain. This will be exploited to design single-channel and multichannel filters with various methods discussed in this book. We also reviewed the main properties of the magnitude spectra of audio sources, that will be used to derive spectral models in the remaining chapters. For advanced readers, the properties of phase spectra and interframe and/or interband filtering techniques are discussed in Chapter 19.

Bibliography

Allen, J. (1977) Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **25** (3), 235 – 238.

Araki, S., Mukai, R., Makino, S., Nishikawa, T., and Saruwatari, H. (2003) The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing*, **11** (2), 109–116.

Avargel, Y. and Cohen, I. (2007) System identification in the short-time Fourier transform domain with crossband filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, **15** (4), 1305 – 1319.

Brown, J. (1991) Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, **89** (1), 425–434.

Burred, J. and Sikora, T. (2006) Comparison of frequency-warped representations for source separation of stereo mixtures, in *Proceedings of the Audio Engineering Society Convention*. Paper number 6924.

Crochiere, R. (1980) A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28** (1), 99 – 102.

Crochiere, R.E. and Rabiner, L.R. (1983) *Multirate Digital Signal Processing*, Prentice Hall.

Duong, N.Q.K., Vincent, E., and Gribonval, R. (2010) Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation, in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 73–80.

Gannot, S., Burshtein, D., and Weinstein, E. (2001) Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, **49** (8), 1614–1626.

Gilloire, A. and Vetterli, M. (1992) Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation. *IEEE Transactions on Signal Processing*, **40** (8), 1862–1875.

Glasberg, B.R. and Moore, B.C.J. (1990) Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, **47**, 103–138.

Gröchenig, K. (2001) *Foundations of Time-Frequency Analysis*, Springer.

Harris, F.J. (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, **66** (1), 51–.83.

ISO (2005) Information technology — Coding of audio-visual objects — Part 3: Audio (ISO/IEC 14496-3:2005).

Kellermann, W. and Buchner, H. (2003) Wideband algorithms versus narrowband

algorithms for adaptive filtering in the DFT domain, in *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, pp. 1278 – 1282.

Makhoul, J. and Cosell, L. (1976) LPCW: An LPC vocoder with linear predictive spectral warping, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*.

Mallat, S. (1999) *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edn..

Mirsamadi, S., Ghaffarzadegan, S., Sheikhzadeh, H., Ahadi, S.M., and Rezaie, A.H. (2012) Efficient frequency domain implementation of noncausal multichannel blind deconvolution for convolutive mixtures of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **20** (8), 2365–2377.

Nesbit, A., Vincent, E., and Plumbley, M.D. (2009) Extension of sparse, adaptive signal decompositions to semi-blind audio source separation, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 605–612.

Portnoff, M.R. (1980) Time-frequency representation of digital signals and systems based on short-time Fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28** (1), 55–69.

Schörkhuber, C. and Klapuri, A. (2010) Constant-Q transform toolbox for music processing, in *Proceedings of Sound and Music Computing Conference*.

Servière, C. (2003) Separation of speech signals with segmentation of the impulse responses under reverberant conditions, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 511–516.

Shynk, J. (1992) Frequency-domain and multirate and adaptive filtering. *IEEE Signal Processing Magazine*, **9** (1), 14–37.

Slaney, M., Naar, D., and Lyon, R.F. (1994) Auditory model inversion for sound separation, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*.

Soo, J.S. and Pang, K.K. (1990) Multidelay block frequency domain adaptive filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38** (2), 373–376.

Stevens, S.S., Volkmann, J., and Newman, E.B. (1937) A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, **8** (3), 185–190.

Vaidyanathan, P.P. (1993) *Multirate Systems And Filter Banks*, Prentice Hall.

Vincent, E. (2006) Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, **14** (1), 91–98.

Vincent, E., Gribonval, R., and Plumbley, M.D. (2007) Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, **87** (8), 1933–1950.

Yılmaz, Ö. and Rickard, S. (2004) Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, **52** (7), 1830–1847.